Large-Scale Genomes Data Inventory of NIH-Funded Research

Erin S. Luetkemeier (OD/OSP), Rob Harriman (OD/DPCPSI), Paula Fearon (OD/DPCPSI), Stephen Sherry (NLM/NCBI), Dina N. Paltoo (OD/OSP)

ABSTRACT

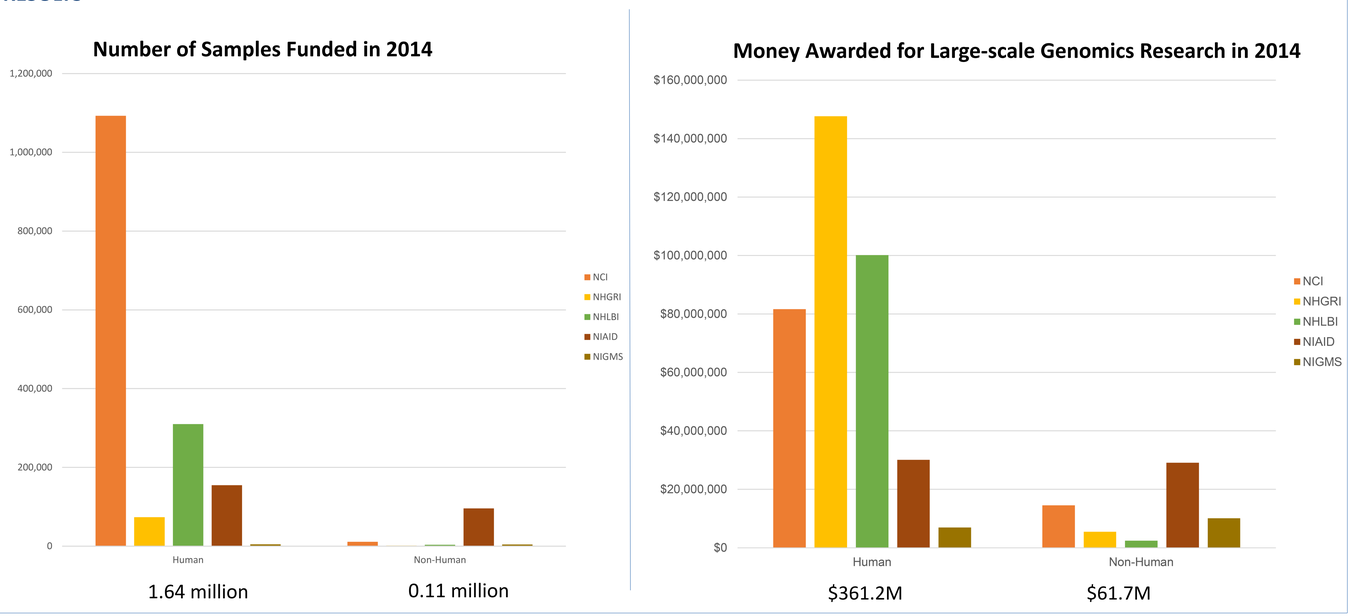
The NIH Genomic Data Sharing (GDS) Policy (http://gds.nih.gov/03policy2.html) was released in August 2014 and applies to all NIH-funded research that generates large-scale human or non-human genomic data. In preparation for the GDS Policy and to determine how much genomic data are likely to be submitted to NIH data repositories, a data inventory was performed in a subset of NIH Institutes and Centers (ICs) with broad funding portfolios in human and non-human genomic studies. The inventory collected the number of studies, number of specimens, funding amounts, and repositories to which the data would likely have been submitted were the studies subject to the Policy. A list of intramural and extramural projects awarded in fiscal year (FY) 2014 for large-scale genomic research was developed by establishing keywords that were relatively unique to the large-scale collection of genomic data such as whole genome sequence, RNAseq, and Bisulfite. The keywords were then used to search titles, abstracts, and specific aims of NIH funded projects. The resulting list of FY 2014 projects potentially within the scope of the GDS Policy was shared with the subset of ICs, who were then asked to use the project list to provide the information requested for the inventory.

The results revealed that, from over 1.75 million samples, the most prominent large-scale genomic data types funded were whole exome sequence, RNAseq, and gene expression, for human studies, and whole genome sequence, RNAseq, and gene expression for non-human studies. These projects resulted in NIH awarding approximately $422M for large-scale genomic research in 2014. From these results, NCBI concluded that it currently has the capacity to meet the expected demands for infrastructure needs under the GDS Policy; however, in order to address future infrastructure needs, NCBI is developing IC costs models to host data from very large human genome sequencing studies.

METHODS

To focus the responses, the subset of ICs were provided a list of applications awarded for large-scale genomic research in 2014. This list was developed by establishing keywords that were relatively unique to the large-scale collection of genomic data which were then used to search titles, abstracts, and specific aims of NIH funded projects. Projects that contained at least one keyword were then subject to the following filters: awarded only, FY: 2014, all mechanisms, all classes (Grants, Intramural, Contracts, IAA's), and HL, AI, CA, GM, and HG. The final list of projects were shared with the subset of ICs and they were asked to supply responses on the number of studies funded in FY 2014 in each large-scale data type; and the total number of individual specimens (human and non-human) from which data have been generated in each data type; the repository to which the data would likely have been submitted were the studies subject to the Policy.

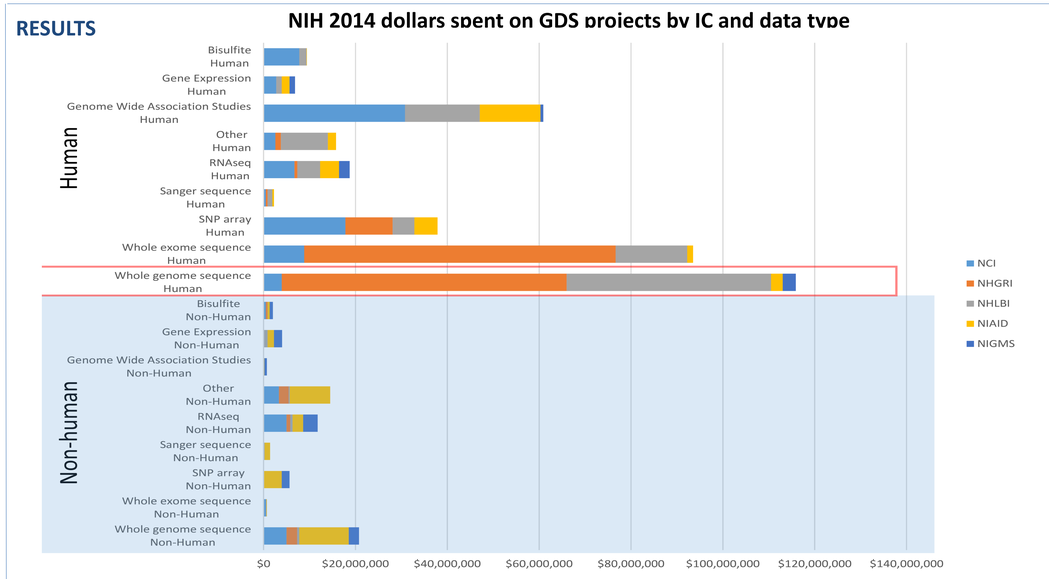
RESULTS

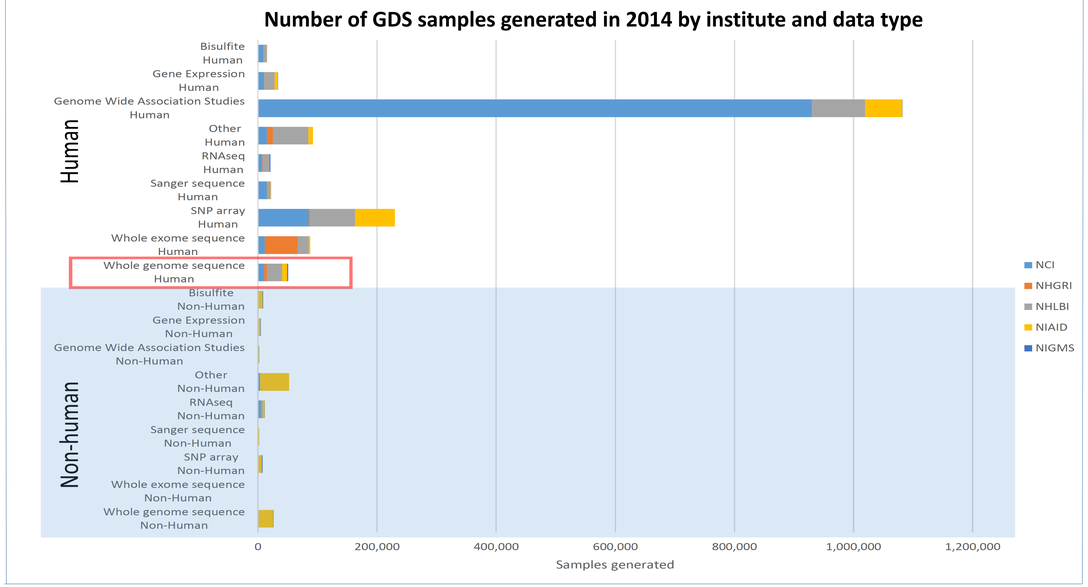


ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of NIH. A special thanks to Sara Dodson (OD/OSP), Marina Volkov (OD/OSP) who contributed to this analysis, and IC representatives from NCI, NHGRI, NHLBI, NIAID, and NIGMS.

RESULTS





CONCLUSIONS/NEXT STEPS

NIH has an opportunity to generate more data per research dollar due to the rapid decline in DNA sequencing costs. This decline, however, has not yet been accompanied by new digital formats that efficiently manage the consequent increase in data volume. Failure to reduce the size of whole genome sequencing studies will lead to unnecessary opportunity costs in terms of access, time, and research dollars. NCBI has identified 'base quality scores' as a sequence property that is both expensive to store and redundant in high coverage sequencing strategies. Our knowledgebase on the value of this property in sequencing projects with deep coverage needs to be extended. Further research is also needed to develop new data formats that efficiently represent the information for large genome sequence projects.